ICONO14

REVISTA ICONO 14, 2009, Nº 13, pp. 334-345. ISSN 1697-8293. Madrid (España)
Basilio Pueo y Victoria Tur: *Sonido espacial para una inmersión audiovisual de alto realismo*
Recibido: 12/07/2009 – Aceptado: 15/10/2009

Nº 13 – REVISTA DE COMUNICACIÓN Y NUEVAS TECNOLOGÍAS – ISSN: 1697 - 8293

SONIDO ESPACIAL PARA UNA INMERSIÓN AUDIOVISUAL DE ALTO REALISMO

Basilio Pueo Ortega

Profesor de Sistemas Audiovisuales

Escuela Politécnica Superior. Universidad de Alicante. Ctra. San Vicente del Raspeig s/n. 03690 Alicante. Tlf: + 34 965903400. Email:basilio@ua.es

Victoria Tur Viñes

Profesora de Creatividad Publicitaria

Facultad Ciencias Económicas y Empresariales. Universidad de Alicante. Ctra. San Vicente del Raspeig s/n. 03690 Alicante. Email: Victoria.Tur@ua.es

Palabras clave

Sonido, Audiovisual, videoconferencia, altavoces

Key Words

Sound, Audiovisual, videoconference, loudspeakers

Abstract

The highly immersive video and audio systems have a major boom in realistic audiovisual environments. The visual and acoustic sensations created in the public are approximated with a high degree of similarity to what is perceived in the real environment intended to be recreated. For that purpose, the stimuli must contain all necessary spatial and temporal information, allowing to create the illusion that the visual object is real. In this article, a review of audiovisual systems that allow this recreation is made, with particular emphasis on surround sound systems. The most promising 3D audio technique, *Wave Field Synthesis*, is described, along with various application fields of highly realistic visual environments.

Resumen

Los sistemas de vídeo y audio de alta inmersión tienen un auge importante en entornos audiovisuales realistas. Las sensaciones visuales y sonoras que crean en el público se aproximan con un alto grado de similitud a lo percibido en el entorno real que pretenden recrear. Para ello, los estímulos deben contener toda la información necesaria, tanto espacial como temporal, que permita crear la ilusión de que el objeto audiovisual es real. En este artículo, se realiza un repaso de los sistemas audiovisuales que permiten esta recreación, con especial atención en los sistemas de audio envolvente. Se describe la técnica de audio 3D más prometedora, Wave Field Synthesis, junto con diversos campos de aplicación de entornos audiovisuales de alto realismo.

Introducción

Los sistemas basados en entornos de telepresencia, como por ejemplo las videoconferencias, se encuentran en el mercado desde hace tiempo. Su objetivo ha sido evitar la necesidad de la presencia física de los asistentes a una reunión. Sin embargo, su impacto en el mercado no ha sido tan importante como se esperaba, debido fundamentalmente, a que la sensación de realismo no era tan elevada como se deseaba. Para mejorar esta sensación de realismo, las investigaciones se encaminan en conseguir que los participantes tengan la sensación de estar físicamente en la reunión.

Objetivos

En este artículo, se presenta el concepto de Ventana Virtual que incluye estímulos visuales y auditivos espaciales, y que constituye el interfaz ideal en videoconferencias de alto realismo. Se presta especial atención a la técnica vanguardista desarrollada en el campo del audio espacial, mediante la introducción de un sistema de inmersión realista y sus aplicaciones en el ámbito cinematográfico, de la videoconferencia y realidad aumentada.

Metodología

Estudio documental de las principales referencias actuales sobre el campo de estudio, y análisis sistemático de los conceptos tratados.

1. Entornos audiovisuales de alto realismo

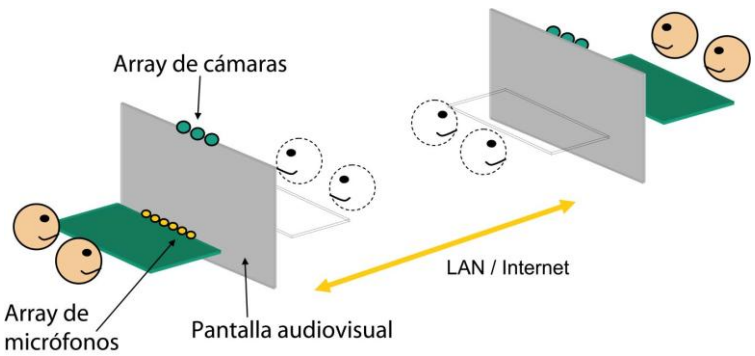
Para conseguir total inmersión, el concepto de ventana virtual debe incluir tanto imagen como sonido de la misma forma que el participante percibiría en una conferencia real. Para ambos estímulos, el sistema de percepción humana puede obtener una sensación tridimensional del espacio usando dos sensores, es decir, dos ojos y dos oídos. La ventana virtual conseguirá por tanto recrear las sensaciones pertinentes si proporciona una percepción realista de la imagen en el espacio en unión con una

sensación de sonido espacial. Dicho concepto está ilustrado en la Figura 1.

Respecto a la imagen, normalmente se emplean dos tecnologías para producir imágenes estereoscópicas: aquellos sistemas en los que el usuario emplea gafas especiales (polarizadas, obturadas o anaglíficas), y por otro lado, los visores autoestereoscópicos que proporcionan percepción tridimensional sin necesidad de llevar gafas especiales o dispositivo alternativo (Dogson 2008).

Respecto al sonido, el método más simple y extendido para proveer sonido espacial es el estéreo, que se ha venido utilizando durante los últimos 50 años como valor añadido de las grabaciones sonoras, sobre todo en la música (Snow 1953). Desde mediados de los años 70 se han venido utilizando en salas de cine inicialmente y en el hogar en los últimos años, los sistemas de sonido envolvente (surround) que intentan proporcionar una mejor sensación que el estéreo utilizando más canales de reproducción (Dolby Surround, Dolby Digital, DTS, SDDS y otros 5.1, 6.1 y 7.1). Sin embargo, estos sistemas sólo tienen como fin incrementar la sensación de espectáculo en las proyecciones cinematográficas añadiendo artificialmente en los procesos de producción, efectos especiales, explosiones, reverberación en altavoces traseros, ambiente, etc., pero no proporcionan una verdadera sensación de sonido 3D. Además la zona útil de escucha (sweet spot) queda prácticamente restringida al punto central del círculo de altavoces, degradándose la percepción fuera del centro. En lo que respecta a la inmersión multimedia realista en videoconferencia, estos sistemas no son adecuados puesto que su objetivo es la

Figura nº 1: Concepto de Ventana Virtual en un sistema multimedia de videoconferencia



Fuente: Propia

reproducción de efectos en películas y los altavoces posteriores no añaden ninguna contribución significativa a la reunión.

Otra estrategia mucho más realista consiste en reproducir directamente en los oídos del oyente la señal que escucharía el oyente si estuviese en el espacio acústico a simular. De la fidelidad de esta reproducción depende la sensación que obtenga dicho oyente. Esta estrategia se denomina comúnmente reproducción de señal binaural y se puede realizar tanto con auriculares como con altavoces (Gardner 1998) Además la señal de sonido 3D puede sintetizarse si se conoce las función HRTF (Head Related Transfer Function) del oyente. Como este sistema es muy sensible a las variaciones de la posición del oyente respecto de la posición óptima de reproducción, en la práctica, sólo son válidos para un único oyente y en entornos de escucha muy controlados, p.e. un usuario delante de la pantalla de un ordenador.

Como alternativa a los sistemas de sonido envolvente existen sistemas más avanzados como Ambisonics o Virtual Surround Panning que son adecuados para zonas de escucha más o menos restringidas (Horbach y Boone, 1999),(Daniel y otros, 1998), aunque siempre algo mayores que los sistemas binaurales con cancelador de cross-talk. La solución para extender la zona de escucha en estos sistemas, implica aumentar el número de altavoces utilizados, con la complejidad y dificultad que implica, así como flexibilizar los formatos de transmisión.

Sin embargo el sistema más prometedor hoy en día para proporcionar una sensación de inmersión sonora en un área muy extensa, es el denominado Wave Field Synthesis (WFS), cuya diferencia fundamental es que el campo sonoro se sintetiza mediante un sistema de arrays de altavoces para toda el área de audiencia, eliminando la zona preferente de escucha. Con WFS se sintetiza el campo acústico que un oyente percibiría en la zona de escucha real, incluido naturalmente todas las colas de localización y efectos que la onda provocaría en el oyente. Esta extensión del sweet spot es muy deseable para aplicaciones multimedia de videoconferencia en tiempo real, en las que múltiples participantes localizados en distintos puntos de una sala interaccionan con otros participantes remotos, también distribuidos espacialmente en una sala. La sensación de realismo se asegura al proveer a los oyentes de la amplitud y dirección de llegada de cualquier mensaje sonoro que les llegue, tanto de los conferenciantes en la misma sala, como los de la sala remota (Brujin y Boone, 2003).

2. Aplicaciones de la técnica

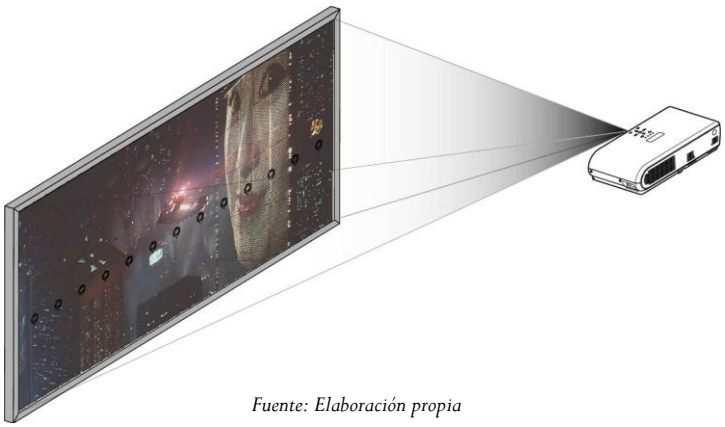
Gracias a los últimos desarrollos y mejoras, se van a poder crear sistemas de teleinmersión para aplicaciones que soporten nuevos servicios interactivos. Dentro del abanico de servicios de telecomunicación, se podrán añadir de manera natural los basados en el estímulo de vídeo y audio volumétrico o estímulo 3D. A continuación, se realizará una breve reseña de las aplicaciones más prometedoras en el campo de los sistemas audiovisuales avanzados.

2.1. Cinematografía envolvente

A pesar de la extensión e importancia de la industria cinematográfica actual, no se ha producido una transición entre los sistemas de cine clásicos y aquellos que incluyan la percepción espacial realista en vídeo y audio. Los pocos sistemas que existen actualmente se encuentran en salas específicas, como auditorios experimentales o museos de ciencia. En estos sistemas, se realiza un gran esfuerzo en dotar de tridimensionalidad a la imagen y se relega el estímulo auditivo a una serie de efectos que simplemente aumentan el dramatismo de la imagen, en lugar de añadir el campo sonoro sintético que estimule nuestros mecanismos de localización de la misma forma que se haría en un entorno real.

Los altavoces convencionales poseen un impacto visual importante que degrada la sensación de inmersión. Además, se necesita el uso de dos arrays de altavoces, encima y debajo de la pantalla para conseguir la

Figura nº 2: Concepto de pantalla MAP con fusión de imagen y audio para una gran audiencia sistema multimedia de videoconferencia



Fuente: Elaboración propia

sensación de que el sonido viene de la propia pantalla. Los prototipos basados en altavoces planos de panel solucionan este problema, ya que permite fusionar altavoz y pantalla en un mismo dispositivo, todo ello de manera transparente al usuario. En la Figura 2 se presenta una ilustración del concepto aquí expuesto.

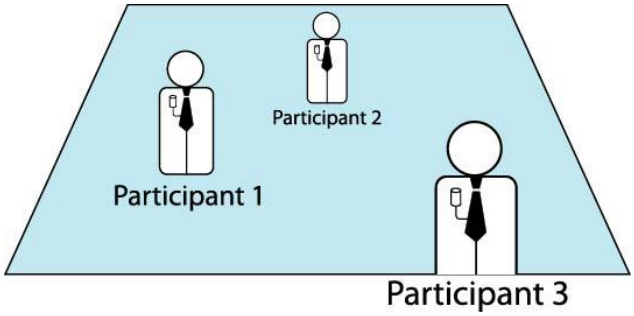
2.2. Teleconferencia audiovisual de alto realismo y resíntesis de WFS

También es posible crear servicios de teleconferencia avanzados con localización de vídeo y audio en tiempo real. Para ello, de entre las diversas tecnologías de visualización 3D basadas en generar dos imágenes en el mismo dispositivo, se propone el uso de la técnica de obturación. Para ello, se emite la señal para el ojo derecho e izquierdo alternativamente al doble de tasa en combinación con unas gafas que bloquean la imagen opuesta adecuadamente. El proyector de imagen debe trabajar a una frecuencia de refresco alta (mayor que 100 Hz) y bajo tecnología DLP. La superficie de proyección es un altavoz plano, el cual funciona como elemento en el que se fusionan los dos estímulos. En la Figura 3 se presenta la disposición de los participantes de la teleconferencia, en la que cada uno de ellos lleva puesto un micrófono de corbata. En esta configuración, los micrófonos se sitúan espacialmente en la sala por medio de un tracking y sus señales, junto con las de posición, se codifican bajo el algoritmo WFS. Si los participantes cambian sus posiciones, el algoritmo actualiza a tiempo real los cambios de modo que en la sala de recepción el sonido siempre provoca la sensación de que viene del orador.

Esta situación es recíproca en la sala de recepción, teniendo los participantes de esa sala la misma disposición de micrófonos y cámara de *tracking*. Toda la información, en ambos sentidos, se codifica, envía y decodifica a tiempo real.

Para facilitar aún más el desarrollo natural

Figura nº 3: Disposición de una teleconferencia con micrófono por participante sistema multimedia de videoconferencia



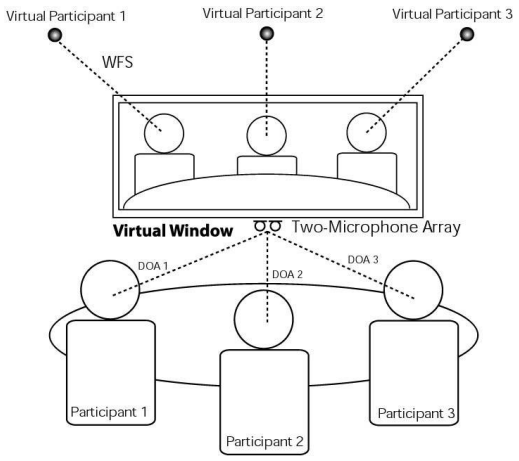
Fuente: Elaboración propia

de la teleconferencia, es posible combinar las ventajas de la fusión de vídeo y audio en el altavoz plano con la automatización en la captación de sonido en los participantes a partir del procesado en array de micrófonos que estiman la dirección de llegada de las señales orales de cada participante. En la Figura 4 se muestra una ilustración del concepto propuesto.

Para desarrollar este concepto, el primer paso es realizar una localización de múltiples oradores, una temática que se ha estudiado intensamente los últimos años debido a sus implicaciones en sistemas de procesado de audio, tales como dispositivos de telefonía móvil. Los algoritmos indirectos de localización de fuentes estiman el retraso temporal de llegada de las señales sonoras entre varios micrófonos y, basado en la geometría del array de micrófonos, estiman las posiciones de las fuentes generadoras por técnicas de optimización (Madhu y Martin, 2008). Para esta aplicación, se propone el uso de tan sólo dos micrófonos por sus ventajas prácticas. No obstante, los modelos a aplicar en una configuración tan sencilla conllevan un procesado de señal específico (Liu y otros, 2000), (Yilmaz y Rickard, 2004).

Una vez que las señales se han capturado y puesto que se tiene que contemplar el caso de que varios participantes hablen a la vez, se debe implementar algoritmos de separación de fuentes (Pedersen y otros, 2007), (Cardoso 1998). Las conversaciones simultáneas entre participantes, si no se separan adecuadamente, llevan asociadas una degradación en la inteligibilidad de las

Figura nº 4: Teleconferencia aplicando el concepto de ventana virtual y algoritmos de dirección de llegada (DOA) para optimizar las comunicaciones orales.



Fuente: Elaboración propia

palabras. La separación basada en segmentación de señales DOA (*Direction Of Arrival*) puede realizarse a partir de enmascaramiento TF (Cobos y López, 2008), inspirada en técnicas de segmentación de imagen (Otsu 1979). Con este procesado, se obtienen las señales de cada uno de los participantes de forma separada y preparada para la próxima etapa.

Finalmente, las señales de voz separadas y la información de posición proveniente del array de dos micrófonos se usan para ajustar las fuentes en el otro lado de la comunicación por medio del algoritmo WFS, lo que se denomina en la literatura técnica *WFS Resynthesis* (Cobos y López, 2009).

Este sistema presenta diversas ventajas con respecto a la teleconferencia clásica:

- **Escalabilidad**: el sistema es independiente del número de participantes en una teleconferencia ya que no hay microfonía dedicada sino un array de sensores que obtiene la dirección de llegada por mediación de procesado de señal.
- **Economía**: No es necesario instalar un micrófono por cada orador y, por tanto, se evita el tratamiento de diversos canales de audio. Si los micrófonos son inalámbricos, el coste del sistema aumenta aún más. Además, no se necesita un sistema de *tracking* para obtener la posición de los participantes.
- **Comodidad**: El sistema es compacto, contiene la pantalla MAP con generación de audio y vídeo inmersivo, y también el array de dos micrófonos comentado anteriormente. Los participantes no tienen porqué instalarse los micrófonos, sino que directamente se sientan y comienzan la sesión. Además, no es necesario apoyo técnico que compruebe la señal de micrófonos en cada participante.

2.2. Realidad aumentada con enriquecimiento de sonido

La realidad aumentada (AR) es un campo de la investigación que trata de la combinación del mundo real y datos generados por ordenador (realidad virtual), donde los objetos gráficos de ordenador se mezclan en escenas reales y en tiempo real (Feiner, 2002). Se trata de una nueva tecnología que aumenta o mejora la visión que el usuario tiene del mundo real con información adicional sintetizada mediante un modelo computerizado. Los usuarios pueden trabajar y examinar objetos 3D reales mientras reciben información adicional sobre estos objetos o sobre la tarea que se está realizando.

La tecnología de altavoz plano puede aplicarse con éxito al desarrollo de un dispositivo de realidad virtual como el ilustrado en la Figura 5(a) con ventajas en la fusión de estímulos importante. La información adicional que muestra el dispositivo se vería enriquecida por señales de audio sincronizadas que cambian sus características a tiempo real en función de la posición en la pantalla del objeto.

Figura nº 5: Concepto de dispositivo de realidad aumentada con enriquecimiento de audio inmersivo

Fuente: Mac Funamizu http://petitinvention.wordpress.com

Fuente: http://www.metroparisiphone.com

Así, el sonido parecería que emerge de los propios objetos, aumentando el grado de verosimilitud de la escena presenciada. Esta característica, que no está presente en los prototipos actuales, tendría futuro no sólo en dispositivos planos como el de la Figura 5(a), sino también en telefonía móvil multimedia, ilustrado en la Figura 5(b). En este caso, la información adicional que el teléfono móvil mostraría en su pantalla, iría acompañada de mensajes sonoros localizados en el espacio gracias a la pantalla con actuadotes piezoeléctricos de pequeño tamaño.

Las aplicaciones en el ámbito de la comunicación comercial son ilimitadas. Encontramos un desarrollo creciente de las aplicaciones de esta tecnología en el PC, con objetivos comerciales. En este caso, en vez de usar el GPS como localizador de la persona y la brújula del móvil para conocer hacia donde mira, se muestra a la cámara web del ordenador un anuncio impreso con un código o una foto y la cámara devuelve una animación en 3D sobre la imagen real. Las marcas exploran cada vez más las posibilidades comerciales de esta nueva tecnología que no resulta excesivamente costosa aunque todavía ofrece poca interacción.

Esta tecnología podría conseguir mayor eficacia en la exposición del público objetivo a los mensajes comerciales, porque la entrega de la información se hace a petición del consumidor que además se muestra interesado, en ese instante, en la marca, producto o servicio. Se trata de una opción altamente interesante para los anunciantes que viven con desconcierto la pérdida de credibilidad y el desperdicio de impactos, en medios tradicionalmente efectivos como la televisión.

La posibilidad de personalización del mensaje comercial gracias a la realidad aumentada, también representa un activo importante. Una vez creado un perfil, en el dispositivo que se utilice, que compile las preferencias o intereses del usuario, los comercios de la zona donde se ubica podrán emitir o no sus mensajes y ofertas, incluso adaptarlos a las peculiaridades del mismo.

Por otro lado, los resultados son medibles en tiempo real y el efecto sorpresa y la sensación de novedad aseguran un recuerdo duradero en el consumidor. Todas estas ventajas nos indican que estamos frente a una nueva forma de comunicación que consigue superar los inconvenientes de los medios tradicionales y evidencia que la comunicación comercial eficaz será factible en un futuro inminente.

Conclusiones

En este artículo se han presentado las técnicas relacionadas con el vídeo y el audio para dotar de sensación de inmersión realista a un entorno audiovisual. Tras una breve introducción histórica acerca de los sistemas de audio envolvente, se ha hecho hincapié en lo referente a la generación de sonido 3D mediante la técnica *Wave Field Synthesis*. De entre las muchas aplicaciones de esta técnica, se han repasado aquellas que mayor impacto tienen en los medios de transmisión audiovisuales, como la cinematografía envolvente, la videoconferencia inmersiva o la realidad aumentada.

Referencias

N. A. Dogson, "Autoesteroscopic 3D displays" Computer Journal, vol. 38, no. 8, pp. 31-36, August 2008

W.B. Snow, "Basic principles of stereophonic sound", Journal of the SMPT, vol. 61 pp. 922-940, 1953.

W.G. Gardner, "3D Audio using loudspeakers", Kluwer Academic Press Editors, Norwell, MA, 1998.

U. Horbach and M. Boone, "Future Transmission and Rendering Formats for Multichannel Sound", 16th AES Conference on Spatial Sound Reproduction, Rovaniemi, Finland, 1999.

J. Daniel et al., "Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions", Proc. 105th AES Conference, San Francisco 1998, preprint 4795.

W. de Brujin and M. Boone "Application of Wave Field Synthesis in life size videoconferencing", in 114th Conv. Audio Eng. Soc., Amsterdam, The Netherlands, Mar 2003.

N. Madhu and R. Martin, "Advances in Digital Speech Transmission", Ed. Wiley-Interscience, 2008.

C. Liu, B. C. Wheeler, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," Journal of the Acoustical Society of America, vol. 108, no. 4, pp. 1888–1905, 2000.

O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Transactions on Signal Processing, vol. 52, no. 7, pp. 1830–1847, July 2004.

S. Pedersen, J. Larsen, U. Kjems, and L. Parra, "Springer Handbook of Speech Processing". Springer Press, 2007, chapter: A Survey of Convolutive Blind Source Separation Methods.

J. F. Cardoso, "Blind signal separation: Statistical principles," in Proccedings of the IEEE, vol. 86, no. 10. IEEE Computer Society Press, October 1998, pp. 2009–2025

M. Cobos and J. J. Lopez, "Stereo audio source separation based on time-frequency masking and multilevel thresholding," Digital Signal Processing, vol. 18, no. 6, pp. 960–976, 2008

N. Otsu, "A threshold selection method from gray-level histogram," IEEE Transactions on System Man Cybernetics, vol. SMC-9, no. 1, pp. 62–66, 1979

M. Cobos and J. J. Lopez, "Resynthesis of wave-field synthesis scenes from stereo mixtures using sound source separation algorithms," Journal of the Audio Engineering Society, accepted for publication, 2009

S. K. Feiner, "Augmented Reality: A New Way of Seeing: Computer scientists are developing systems that can enhance and enrich a user's view of the world". Scientific American, April 2002.

Cita de este artículo
Pueo, B. y Tur V. (2010) Sonido espacial para una inmersión audiovisual de alto realismo *Revista Icono14 [en línea] 15 de Octubre de 2009, Nº 13.* pp. 334-345. Recuperado (Fecha de acceso), de http://www.icono14.net